University of Zurich[UZH]

Zurich Open Repository and Archive
University of Zurich
University Library
Strickhofstrasse 39
CH-8057 Zurich
www.zora.uzh.ch

Year: 2011

Hyper-g priors for generalised additive model selection

Sabanés Bové, D ; Held, L ; Kauermann, G

Abstract: We propose an automatic Bayesian approach to the selection of covariates and penalised splines transformations thereof in generalised additive models. Specification of a hyper-g prior for the model parameters and a multiplicity-correction prior for the models themselves is crucial for this task. We introduce the methodology in the normal model and illustrate it with an application to diabetes data. Extension to non-normal exponential families is finally discussed.

Posted at the Zurich Open Repository and Archive, University of Zurich
ZORA URL: https://doi.org/10.5167/uzh-53039
Conference or Workshop Item
Accepted Version

Originally published at:
Sabanés Bové, D; Held, L; Kauermann, G (2011). Hyper-g priors for generalised additive model selection. In: 26th International Workshop on Statistical Modelling, Valencia, Spain, 10 July 2011 - 15 July 2011, 538-543.

Student Oral Presentation

Hyper-g Priors for Generalised Additive Model Selection

Daniel Sabanés Bové[1], Leonhard Held[1] and Göran Kauermann[2]

[1] Division of Biostatistics, Institute of Social and Preventive Medicine, University of Zurich, Hirschengraben 84, CH-8001 Zurich, Switzerland, Email: daniel.sabanesbove@ifspm.uzh.ch, leonhard.held@ifspm.uzh.ch

[2] Centre for Statistics, Department of Economics and Business Administration, University Bielefeld, Postfach 300131, D-33501 Bielefeld, Germany, Email: gkauermann@uni-bielefeld.de

Abstract: We propose an automatic Bayesian approach to the selection of covariates and penalised splines transformations thereof in generalised additive models. Specification of a hyper-g prior for the model parameters and a multiplicity-correction prior for the models themselves is crucial for this task. We introduce the methodology in the normal model and illustrate it with an application to diabetes data. Extension to non-normal exponential families is finally discussed.

Keywords: Penalised splines; Bayesian variable selection; Shrinkage.

1 Introduction

Suppose we have p metrical covariates $x_1, \ldots, x_p$ and use the additive model $y = \beta_0 + \sum_{j=1}^{p} m_j(x_j) + \epsilon$, where $\epsilon \sim N(0, \sigma^2)$. When x_j is included non-linearly in the model, we assume

$$m_j(x_j) = x_j\beta_j + \boldsymbol{Z}_j(x_j)^T\boldsymbol{u}_j$$

where $\boldsymbol{Z}_j(x_j)$ is the $K \times 1$ spline basis vector at position x_j and $\boldsymbol{u}_j \sim N(\boldsymbol{0}, \sigma^2\rho_j\boldsymbol{I})$ is the corresponding coefficients vector. In order to combine n observations, we stack these to the $n \times 1$ vector $\boldsymbol{x}_j$ and the $n \times K$ basis matrix $\boldsymbol{Z}_j$, both modified to be zero-centred and orthogonal to each other. We then translate the variance parameter ρ_j into the corresponding degree of freedom (Aerts, Claeskens and Wand, 2002, section 2.2)

$$d_j(\rho_j) = \mathrm{tr}\{(\boldsymbol{Z}_j^T\boldsymbol{Z}_j + \rho_j^{-1}\boldsymbol{I})^{-1}\boldsymbol{Z}_j^T\boldsymbol{Z}_j\} + 1 \in (1, K+1). \qquad (1)$$

A larger ρ_j (or a larger d_j) leads to a weaker penalty on the non-linear component of the function m_j. If x_j is excluded from or linearly included in the model we have $m_j(x_j) \equiv 0$ or $m_j(x_j) = x_j\beta_j$ and set $d_j = 0$ or $d_j = 1$, respectively. Thus, the function m_j is exactly defined by d_j, which we may restrict to a finite set of values, say $d_j \in \{0, 1\} \cup \{2, 3, \ldots, K\}$.

As default prior for the parameters β_0, $\boldsymbol{\beta} = (\beta_j : d_j \geq 1)$ and σ^2 in a given model specified via $\boldsymbol{d} = (d_1, \ldots, d_p)$,

$$\boldsymbol{y} \,|\, \beta_0, \boldsymbol{\beta}, \boldsymbol{u}, \sigma^2 \sim \mathrm{N}(\mathbf{1}\beta_0 + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u},\, \sigma^2 \boldsymbol{I}) \tag{2}$$

with $\boldsymbol{X} = (\boldsymbol{x}_j : d_j \geq 1)$, $\boldsymbol{Z} = (\boldsymbol{Z}_j : d_j > 1)$ and $\boldsymbol{u} = (\boldsymbol{u}_j^T : d_j > 1)^T$, we propose the hyper-$g$ prior (Liang *et al.*, 2008) described in Section 2. For the models we propose a multiplicity-correction prior in Section 3. The methodology is applied to diabetes data in Section 4 and extended to generalised additive models in Section 5.

2 Hyper-g Priors for Additive Models

Integrating out the spline coefficients vector $\boldsymbol{u} \sim \mathrm{N}(\mathbf{0}, \sigma^2 \boldsymbol{D})$, where $\boldsymbol{D} = \mathrm{diag}\{\rho_j \boldsymbol{I} : d_j > 1\}$, from the conditional model (2) yields the marginal model

$$\boldsymbol{y} \,|\, \beta_0, \boldsymbol{\beta}, \sigma^2 \sim \mathrm{N}\left(\mathbf{1}\beta_0 + \boldsymbol{X}\boldsymbol{\beta},\, \sigma^2 \boldsymbol{V}\right) \tag{3}$$

with $\boldsymbol{V} = \boldsymbol{I} + \boldsymbol{Z}\boldsymbol{D}\boldsymbol{Z}^T$ having Cholesky decomposition $\boldsymbol{V} = \boldsymbol{R}^T\boldsymbol{R}$. The transformed response vector $\tilde{\boldsymbol{y}} = \boldsymbol{R}^{-T}\boldsymbol{y}$ follows a linear model with similarly transformed design matrix $\tilde{\boldsymbol{X}}$ and diagonal covariance matrix $\sigma^2\boldsymbol{I}$. It turns out that we can use the hyper-g prior (Liang *et al.*, 2008) for this transformed model, *i. e.* a locally uniform prior $p(\beta_0) \propto 1$ on the intercept, Jeffreys' prior $p(\sigma^2) \propto (\sigma^2)^{-1}$ on the variance and the g-prior (Zellner, 1986)

$$\boldsymbol{\beta} \,|\, g, \sigma^2 \sim \mathrm{N}\left(\mathbf{0},\, g\sigma^2(\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}})^{-1}\right) \tag{4}$$

on the coefficients are combined with a uniform prior on the shrinkage coefficient $g/(1+g)$. Note that $\sigma^{-2}\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}} = \sigma^{-2}\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X}$ is the Fisher information matrix of $\boldsymbol{\beta}$ in the marginal model (3). The hyper-g prior leads to a closed form of the marginal likelihood, which we can compute on the original response scale via the change of variables formula:

$$p(\boldsymbol{y} \,|\, \boldsymbol{d}) \propto \|\tilde{\boldsymbol{y}} - \bar{\tilde{\boldsymbol{y}}}\|^{-(n-1)} \, (l_{\boldsymbol{d}} + 2)^{-1} \, {}_2F_1\left(\frac{n-1}{2}; 1; \frac{l_{\boldsymbol{d}} + 4}{2}; \tilde{R}^2\right) |\boldsymbol{R}|^{-1},$$

where $l_{\boldsymbol{d}}$ is the dimension of $\boldsymbol{\beta}$, ${}_2F_1$ is the Gaussian hypergeometric function and $\tilde{R}^2$ is the classical coefficient of determination in model (3).

3 Model Prior

We propose a prior $p(\boldsymbol{d})$ on the model space which explicitly corrects for the multiplicity of testing inherent in the simultaneous analysis of many covariates (see Scott and Berger, 2010): *A priori*, the number of covariates included in the model ($l_{\boldsymbol{d}}$) is uniformly distributed on $\{0, 1, \ldots, p\}$.

TABLE 1. Marginal posterior probabilities (x_1: age, x_2: systolic blood pressure, x_3: cholesterol/HLD ratio, x_4: BMI, x_5: waist/hip ratio, x_6: gender).

	x_1	x_2	x_3	x_4	x_5	x_6
not included ($d_j = 0$)	0.00	0.65	0.00	0.14	0.50	0.65
linear ($d_j = 1$)	0.71	0.33	0.93	0.81	0.48	0.35
non-linear ($d_j > 1$)	0.29	0.03	0.07	0.05	0.02	—

Then the number of non-linearly included covariates ($s_{\boldsymbol{d}}$) is uniformly distributed on $\{0, 1, \ldots, l_{\boldsymbol{d}}\}$. The respective choice of the $l_{\boldsymbol{d}}$ and $s_{\boldsymbol{d}}$ covariates is uniformly distributed on all possible configurations. Finally, the degrees of freedom of the non-linearly modelled covariates are independent and uniformly distributed on $\{2, 3, \ldots, K\}$. Altogether, this gives

$$1/p(\boldsymbol{d}) = \binom{p}{l_{\boldsymbol{d}}}(p+1)\binom{l_{\boldsymbol{d}}}{s_{\boldsymbol{d}}}(l_{\boldsymbol{d}}+1)(K-1)^{s_{\boldsymbol{d}}}$$

and leads to marginal prior probabilities $\Pr(d_j = 0) = 1/2$, $\Pr(d_j = 1) = \Pr(d_j > 1) = 1/4$.

4 Application

We illustrate our modelling approach with the diabetes data from Harrell (2001). We study the association of (the negative reciprocal of) glycosolated haemoglobin of $n = 377$ study participants with the continuous covariates age (in years), systolic blood pressure (in mmHg), cholesterol/HDL ratio, body mass index (BMI, in kg/m^2) and waist/hip ratio as well as the binary covariate gender. As the computational complexity is quadratic in the spline basis dimension K, we want to use splines with few quantile-based knots. Therefore, we choose cubic O'Sullivan splines (Wand and Ormerod, 2008). Here, we get basis matrices $\boldsymbol{Z}_j$ with $K = 9$ columns from 7 knots. The exhaustive evaluation of the posterior model probabilities $p(\boldsymbol{d} \mid \boldsymbol{y}) \propto p(\boldsymbol{y} \mid \boldsymbol{d})p(\boldsymbol{d})$ of all $(K+1)^5 \cdot 2 = 200\,000$ models takes only 585 seconds due to an efficient `C++` implementation which is available in an `R`-package from the first author. In Table 1 the marginal posterior probabilities for linear and non-linear inclusion of the six covariates are shown. There is strong evidence for linear inclusion of cholesterol/HDL ratio and BMI, while the posterior probability for inclusion of systolic blood pressure or gender is only 35%. There is overwhelming evidence for (non-linear) inclusion of age, and the posterior odds for (linear) inclusion of waist/hip ratio are around 1. The *maximum a posteriori* model includes age, cholesterol/HDL ratio and BMI all linearly. Note that these are the covariates which have inclusion probabilities larger than 50%, thus defining the set of median probability models (Barbieri and Berger, 2004) $\boldsymbol{d}$ with $d_1, d_3, d_4 \geq 1$

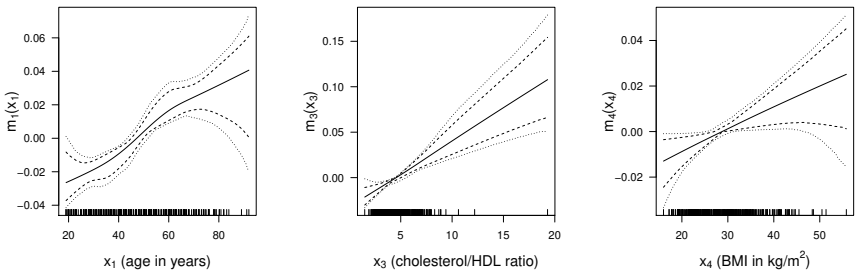

FIGURE 1. Estimated covariate effects in the median probability model average, based on 10 000 samples: Posterior means (solid lines), pointwise (dashed lines) and simultaneous (dotted lines) 95%-credible intervals as well as positions of data points (ticks above x-axes) are shown.

and $d_2 = d_5 = d_6 = 0$. Figure 1 shows the estimated covariate effects from the resulting model average. While the age effect is slightly non-linear (with 38% probability in the median probability models), both other covariates have essentially linear effect estimates.

5 Extension to Generalised Additive Models

Now we assume more generally that the covariate effects $m_j(x_j)$ enter additively into the linear predictor $\eta = \beta_0 + \sum_{j=1}^{p} m_j(x_j)$ of an exponential family distribution with canonical parameter θ, mean $\mathrm{E}(y) = h(\eta) = db(\theta)/d\theta$ and variance $\mathrm{Var}(y) = \phi/w \cdot v(\mu) = \phi/w \cdot d^2b(\theta)/d\theta^2$ (see McCullagh and Nelder, 1989). We restrict our attention to non-normal distributions with fixed dispersion ϕ (as $\phi = 1$ for the Bernoulli and Poisson distribution) and known weight w. For n observations, the linear predictor vector $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)^T$ is $\boldsymbol{\eta} = \mathbf{1}\beta_0 + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u}$, where $\boldsymbol{u} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{D})$, and the likelihood is

$$p(\boldsymbol{y} \mid \beta_0, \boldsymbol{\beta}, \boldsymbol{u}) \propto \exp\left\{ \sum_{i=1}^{n} \frac{y_i\theta_i - b(\theta_i)}{\phi/w_i} \right\}. \tag{5}$$

A reasonable generalisation of (1) is (see Ruppert, Wand and Carroll, 2009, section 11.4)

$$d_j(\rho_j) = \mathrm{tr}\{(\boldsymbol{Z}_j^T \widehat{\boldsymbol{W}} \boldsymbol{Z}_j + \rho_j^{-1}\boldsymbol{I})^{-1} \boldsymbol{Z}_j^T \widehat{\boldsymbol{W}} \boldsymbol{Z}_j\} + 1, \tag{6}$$

which uses a fixed weight matrix $\widehat{\boldsymbol{W}} = \boldsymbol{W}(\mathbf{1}\widehat{\beta}_0)$ for all models, where $\boldsymbol{W}(\boldsymbol{\eta}) = \mathrm{diag}\{(dh(\eta_i)/d\eta)^2 v(h(\eta_i))^{-1}\phi^{-1}w_i\}_{i=1}^n$ is the usual generalised linear model (GLM) weight matrix and $\widehat{\beta}_0$ is the intercept estimate from the null model. Therefore, we now arrange $\mathbf{1}$, $\boldsymbol{x}_j$ and the columns of $\boldsymbol{Z}_j$ to be orthogonal with respect to the inner product in terms of $\widehat{\boldsymbol{W}}$, so that (6)

correctly captures the degrees of freedom associated with the non-linear part of m_j.

In order to derive a generalised g-prior for $\boldsymbol{\beta}$, we will use the iterative weighted least squares (IWLS) approximation to (5) to come back to a normal model and then derive the resulting g-prior (4). So let

$$\boldsymbol{z}_0 = \boldsymbol{\eta}_0 + \mathrm{diag}\{dh(\boldsymbol{\eta}_0)/d\boldsymbol{\eta}\}^{-1}(\boldsymbol{y} - h(\boldsymbol{\eta}_0))$$

be the adjusted response vector resulting from a first-order approximation to $h^{-1}(\boldsymbol{y})$ around $h(\boldsymbol{\eta}_0)$, such that

$$\boldsymbol{z}_0 \,|\, \beta_0, \boldsymbol{\beta}, \boldsymbol{u} \sim \mathrm{N}\big(\mathbf{1}\beta_0 + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u},\, \boldsymbol{W}(\boldsymbol{\eta}_0)^{-1}\big)$$

is the working normal model. This can be rewritten to

$$\tilde{\boldsymbol{z}}_0 \,|\, \beta_0, \boldsymbol{\beta}, \boldsymbol{u} \sim \mathrm{N}(\tilde{\mathbf{1}}\beta_0 + \tilde{\boldsymbol{X}}\boldsymbol{\beta} + \tilde{\boldsymbol{Z}}\boldsymbol{u},\, \boldsymbol{I}) \tag{7}$$

by setting $\tilde{\boldsymbol{z}}_0 = \boldsymbol{W}(\boldsymbol{\eta}_0)^{1/2}\boldsymbol{z}_0$ *etc.* Since (7) is analogous to (2), our proposal for a generalised g-prior is

$$\boldsymbol{\beta} \,|\, g \sim \mathrm{N}(\mathbf{0}, g\boldsymbol{J}^{-1}), \tag{8}$$

where $\boldsymbol{J}$ is the Fisher information for $\boldsymbol{\beta}$ in (7) with $\boldsymbol{\eta}_0 = \mathbf{0}$:

$$\begin{aligned}\boldsymbol{J} &= \tilde{\boldsymbol{X}}^T(\boldsymbol{I} + \tilde{\boldsymbol{Z}}\boldsymbol{D}\tilde{\boldsymbol{Z}}^T)^{-1}\tilde{\boldsymbol{X}} \\ &= \boldsymbol{X}^T\boldsymbol{W}_0^{1/2}(\boldsymbol{I} + \boldsymbol{W}_0^{1/2}\boldsymbol{Z}\boldsymbol{D}\boldsymbol{Z}^T\boldsymbol{W}_0^{1/2})^{-1}\boldsymbol{W}_0^{1/2}\boldsymbol{X},\end{aligned}$$

abbreviating $\boldsymbol{W}_0 = \boldsymbol{W}(\mathbf{0})$. Note that this prior directly generalises the prior proposed by Sabanés Bové and Held (2011) for GLMs, to which it reduces when there are no spline effects in the model.

The generalised hyper-g prior then consists of the improper prior $p(\beta_0) \propto 1$ on the intercept β_0, the g-prior (8) on the linear effects vector $\boldsymbol{\beta}$, the penalty prior $\boldsymbol{u} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{D})$ on the spline coefficients vector $\boldsymbol{u}$ and some proper hyper-prior $p(g)$ on the hyper-parameter g in the g-prior. For the implementation of posterior inference we can easily extend the approach of Sabanés Bové and Held (2011, section 3). Let $\boldsymbol{X}_a = (\mathbf{1}, \boldsymbol{X}, \boldsymbol{Z})$ and $\boldsymbol{\beta}_a = (\beta_0, \boldsymbol{\beta}^T, \boldsymbol{u}^T)^T$, such that $\boldsymbol{\eta} = \boldsymbol{X}_a\boldsymbol{\beta}_a$. The prior for $\boldsymbol{\beta}_a$ conditional on g has Gaussian form with mean zero and singular precision $\boldsymbol{R}_a = \mathrm{diag}\{0, g^{-1}\boldsymbol{J}(\mathbf{0}), \boldsymbol{D}^{-1}\}$. Thus, the Laplace approximation of $p(\boldsymbol{y} \,|\, g, \boldsymbol{d})$, which is based on a Gaussian approximation to the conditional posterior $p(\boldsymbol{\beta}_a \,|\, \boldsymbol{y}, g)$, can be obtained by the Bayesian IWLS algorithm (West, 1985). Afterwards, the marginal likelihood

$$p(\boldsymbol{y} \,|\, \boldsymbol{d}) = \int_0^\infty p(\boldsymbol{y} \,|\, g, \boldsymbol{d})p(g)\, dg,$$

can be approximated by numerical integration of the Laplace approximation $\tilde{p}(\boldsymbol{y} \,|\, g, \boldsymbol{d})$. Note that this strategy of integrated Laplace approximations was proposed more generally by Rue, Martino and Chopin (2009). Finally, for sampling from the posterior of $\boldsymbol{\beta}_a$ and g in a specific model $\boldsymbol{d}$ we can use a tuning-free Metropolis-Hastings algorithm.

References

Aerts, M., Claeskens, G., and Wand, M.P. (2002). Some theory for penalized spline generalized additive models. *Journal of Statistical Planning and Inference*, **103**, 455-470.

Barbieri, M.M., and Berger, J.O. (2004). Optimal predictive model selection. *Annals of Statistics*, **32**, 870-897.

Harrell, Jr., F.E. (2001). *Regression Modeling Strategies*. New York: Springer.

Liang, F., Paulo, R., Molina, G., Clyde, M.A., and Berger, J.O. (2008). Mixtures of g priors for Bayesian variable selection. *Journal of the American Statistical Association*, **103**, 410-423.

McCullagh, P., and Nelder, J.A. (1989). *Generalized Linear Models*. New York: Chapman and Hall.

Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society. Series B (Methodological)*, **71**, 319-392.

Ruppert, D., Wand, M.P., and Carroll, R.J. (2003). *Semiparametric Regression*. Cambridge: Cambridge University Press.

Sabanés Bové, D., and Held, L. (2011). Hyper-g Priors for Generalized Linear Models. *Bayesian Analysis*, **6**, forthcoming article. URL: `http://ba.stat.cmu.edu/abstracts/Sabanes.php`

Scott, J.G., and Berger, J.O. (2010). Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *Annals of Statistics*, **38**, 2587-2619.

Wand, M.P., and Ormerod, J.T. (2008). On semiparametric regression with O'Sullivan penalized splines. *Australian & New Zealand Journal of Statistics*, **50**, 179-198.

West, M. (1985). Generalized linear models: scale parameters, outlier accommodation and prior distributions. In: *Bayesian Statistics 2: Proceedings of the Second Valencia International Meeting*, 531-558. Amsterdam: North-Holland.

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In: *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*, 233-243. Amsterdam: North-Holland.